Version RC/4

EGT3
ENGINEERING TRIPOS PART IIB

Wednesday 7 May 2025 2 to 3.40

Module 4F12

COMPUTER VISION

*Answer not more than **three** questions.*

All questions carry the same number of marks.

*The **approximate** percentage of marks allocated to each part of a question is indicated in the right margin.*

*Write your candidate number **not** your name on the cover sheet.*

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed
Engineering Data Book

10 minutes reading time is allowed for this paper at the start of the exam.

You may not start to read the questions printed on the subsequent pages of this question paper until instructed to do so.

You may not remove any stationery from the Examination Room.

1 In computer vision, point correspondences over different viewpoints are often used to recover an object's pose and 3-D shape. Point features are first detected in each image and then matched to features in the other viewpoints.

(a) The detection of features first requires the smoothing of a greyscale image, $I(x, y)$, by *low-pass* filtering with a Gaussian filter, $G_\sigma(x, y)$, at different scales σ.

(i) What are the properties of a Gaussian kernel that make it a suitable low-pass filter? [10%]

(ii) How is low-pass filtering at multiple scales implemented efficiently using an *image pyramid*? [15%]

(iii) Describe the convolutions required to construct an image pyramid of low-pass filtered images, $S_{\sigma_0} \ldots S_{8\sigma_0}$, with $s = 3$ distinct images in each octave. [15%]

(b) An image feature is usually localised in position and scale by filtering the image with a *band-pass* filter over different scales.

(i) Describe a suitable band-pass filter that can be used for finding blobs in images. [10%]

(ii) How are the blob centres localised by this filter? Show how to compute the blob's apparent size. [10%]

(c) The SIFT (Scale-Invariant Feature Transform) descriptor is used to describe each feature. It is computed from a 16×16 patch of pixels around each feature centre.

(i) How is the 16×16 patch of pixels sampled at an appropriate scale and orientation from the image pyramid? [15%]

(ii) How does the SIFT descriptor achieve its invariance to viewpoint changes, lighting and occlusion? [15%]

(iii) What are its limitations? [10%]

2 A camera is used to view a three-dimensional object under perspective projection such that the image co-ordinates, (u_i, v_i), of a 3-D world point, (X_i, Y_i, Z_i), can be modelled by:

$$u_i = \frac{p_{11}X_i + p_{12}Y_i + p_{13}Z_i + p_{14}}{p_{31}X_i + p_{32}Y_i + p_{33}Z_i + p_{34}}$$

$$v_i = \frac{p_{21}X_i + p_{22}Y_i + p_{23}Z_i + p_{24}}{p_{31}X_i + p_{32}Y_i + p_{33}Z_i + p_{34}}$$

(a) (i) This mapping from 3-D world to image co-ordinates can also be expressed as a 3×4 projection matrix. Give a factorisation of the projection matrix into components which encode the camera position, camera orientation and internal camera calibration parameters. [10%]

(ii) Under what viewing conditions is this model valid and state any assumptions? [10%]

(iii) Find the *vanishing* point of lines which are parallel to the Z-axis. [5%]

(iv) What is meant by the *horizon*? Derive the equation of the horizon of the $X - Y$ ground plane. [15%]

(b) The camera is to be calibrated by viewing a calibration object with a set of N known reference 3-D points, $\{(X_i, Y_i, Z_i)\}_{i=1}^{N}$, and measuring the corresponding projections, $\{(u_i, v_i)\}_{i=1}^{N}$.

(i) What is meant by camera calibration? List the parameters that need to be estimated. [10%]

(ii) Show how to calibrate the camera. Give details of the equations and optimisation techniques that are needed when there are a large number of 3-D reference points and the image measurements are noisy. [20%]

(c) The camera is used to view a circle on the $X - Y$ plane.

(i) Derive the equation of the perspective projection of the circle. [20%]

(ii) How would the projection and shape differ under *weak perspective*? [10%]

3 An object is viewed from two viewpoints with a mobile phone camera. The correspondences in the left and right images, (u, v) and (u', v'), satisfy the *epipolar* constraint:

$$\begin{bmatrix} u' & v' & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = 0$$

(a) (i) How is this constraint used differently in stereo vision compared to structure from motion? [10%]

(ii) By considering the relative position and orientation of the camera in the two positions, derive the epipolar constraint. [20%]

(iii) Give an algebraic expression for the epipolar line corresponding to a point in the left image with pixel co-ordinates (u, v) and show how to compute the epipole in the right image. [10%]

(b) The *fundamental matrix* is to be estimated from point correspondences.

(i) First a large number of *keypoints* are detected in each image and potential matches between the two images are found by comparing their descriptors. Describe two possible keypoint descriptors and give an algorithm for determining matches between the views. [15%]

(ii) How are consistent matches obtained in the presence of incorrect or outlier measurements? How is the fundamental matrix estimated when a large number of consistent matches is available? What additional constraint needs to be enforced? [15%]

(c) The *fundamental matrix* can be decomposed to recover the projection matrices for the two viewpoints.

(i) Describe an algorithm to recover the camera motion (rotation and translation) between the two views and the projection matrices if the internal camera parameters (represented by the matrix K) are known. Include details of how any ambiguities are resolved. [20%]

(ii) Show how to recover the 3-D positions of the keypoints. [10%]

4 (a) A typical Convolutional Neural Network (CNN) incorporates convolutional and pooling layers, connected via non-linear activations. Consider a single convolutional layer l in an CNN. Let $\mathbf{I}_l$ be a $C \times N \times N$ input image to the layer, with C channels and spatial dimension N. This layer has K $C \times 3 \times 3$ filters $\mathbf{F}_{lk}$, and produces an output image $\mathbf{O}_l$ of size $K \times N \times N$. Assume zero padding wherever needed.

(i) Describe the forward pass of the layer mathematically. [10%]

(ii) This CNN has a loss function $\mathcal{L}$, and the partial derivatives $\frac{\partial \mathcal{L}}{\partial \mathbf{O}_l}$ are known. Derive an expression for the partial derivatives $\frac{\partial \mathcal{L}}{\partial \mathbf{F}_{lk}}$. [20%]

(iii) Derive an expression for the partial derivatives $\frac{\partial \mathcal{L}}{\partial \mathbf{I}_l}$. [20%]

(b) You are asked to train a Deep Neural Network (DNN) to perform heart-rate estimation from videos of human faces. The network will need to predict a single scalar value correlated to the level of oxygenated blood in the face. A small, paid study is conducted resulting in a database of RGB videos correlated with output from a photoplethysmogram (PPG *i.e.*, blood oxygenation measurements) for use in training the DNN. The period of the per-frame DNN output will be used to calculate the heart rate.

(i) What best practices can you apply to the dataset to increase the likelihood of the DNN generalising well to unseen data? Why might it be a good idea to normalise the values from the PPG to range from -1 to 1? [10%]

(ii) When you append the first image from each video to subsequent images and attempt to predict the PPG difference (as opposed to the absolute value), the DNN performs better. Why might this be the case? [5%]

(c) A larger and more diverse dataset is gathered, so you decide to train the DNN as part of a Transformer architecture.

(i) Why is a Transformer architecture a good choice for this problem? [5%]

(ii) Sketch a Transformer architecture for this problem (with labels), showing where your DNN would fit in the larger system. Your diagram should contain an Encoder and a Decoder, and should show how the repeating block structures of each interact via attention mechanisms. [20%]

(iii) Applying the model at test-time requires decoding the output of the Transformer one frame at a time, dependent on the previous frames. How can you design the network to express its uncertainty in its predictions? What is beam search, and how can it be used to improve the quality of the final output? [10%]

END OF PAPER

THIS PAGE IS BLANK